The application of hierarchical cluster analysis to the selection of isomorphic crystals

Rita Giordano

The effects of radiation damage in macromolecular crystallography cannot be overcome. The situation can only be improved by merging data from a subset of homogeneous samples. This approach can be applied to assemble a complete data set and to increase the signal to noise level of experimental data. The anomalous signal, success in SAD substructure determination and accuracy of phases and electron-density maps all improve with an increase in the number of crystals used in merging.

we evaluated the possibility of using hierarchical cluster analysis using the cluster distance metric based on the intensity correlation coefficients as an instrument for such a selection.

SAD DATA COLLECTION

Type of measurement	Total number of crystals	Inverse beam strategy	X-ray Energy (keV)	Resolution (Å)	Rotation range (º)	Oscillation range (º)	Dose (MGy)
Insulin	12	Yes	9	2.2	50 + 50	2	1
Trypsin	17	No	9	2.2	360	2.25	1
AroF	12	Yes	12.6	3.5	90+90	1.5	1
Thaumatin	13	Yes	6.8	2	100+100	1	2

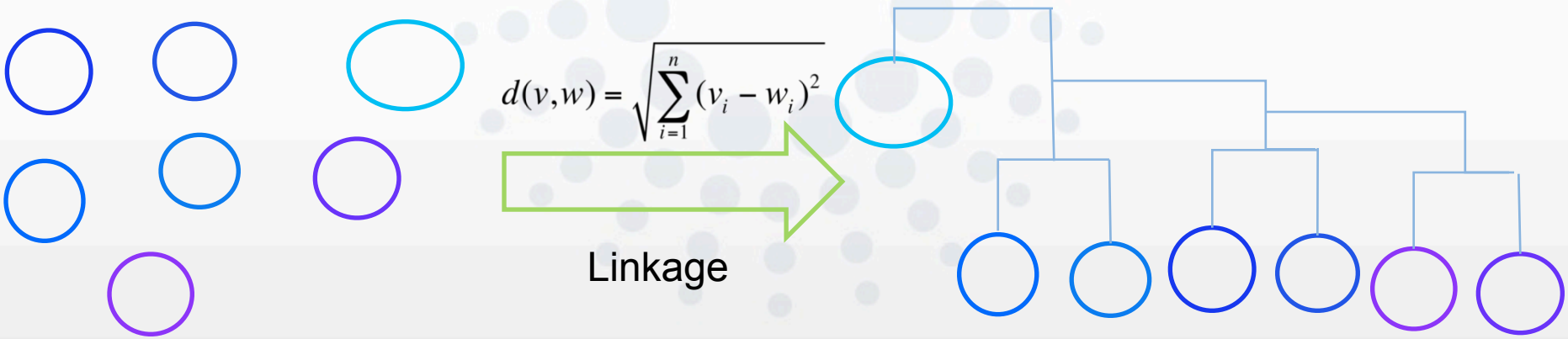
Insulin : 3 loops
- Loop1, 8 crystals
- Loop2, 2 crystals
- Loop3, 3 crystals

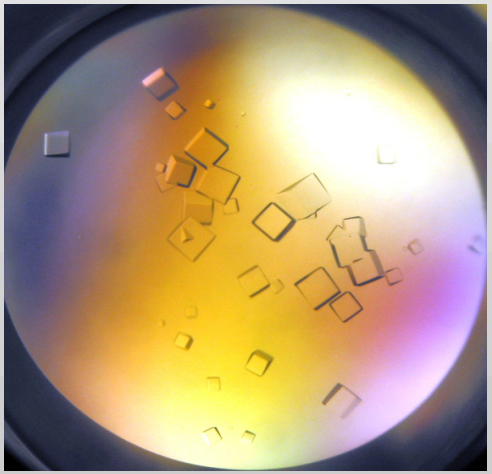
AroF : 3 loops
- Loop1, 3 crystals
- Loop2, 4 crystals
- Loop3, 5 crystals

Trypsin : 15 loops
- Loop1, 6 crystals
- Loop2, 2 crystals
- Loop3, 3 crystals
- 11crystals mounted individually

Thaumatin: loops 9
- Loop1, 6 crystals
- Loop2, 2 crystals
- 7 crystals mounted individually

Hierarchical Cluster Analysis

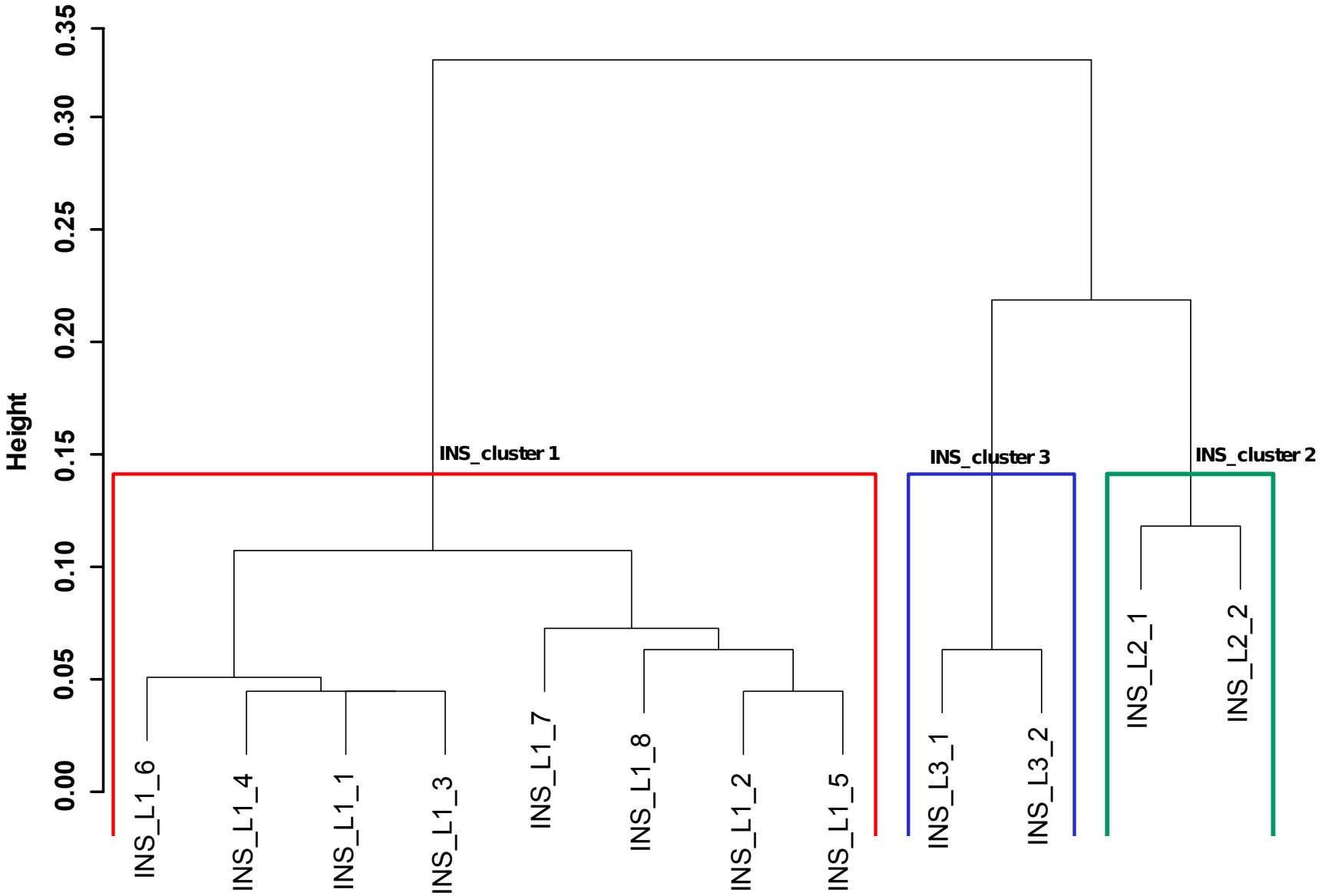




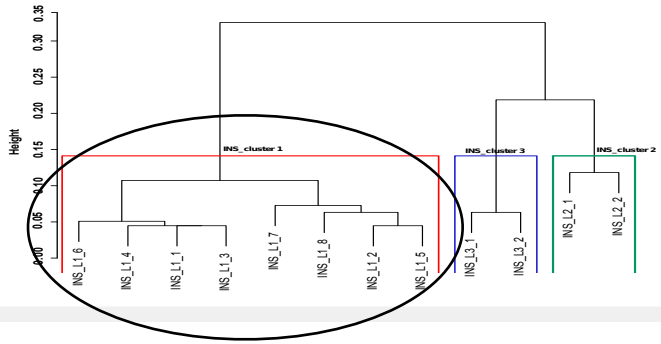
$d(i,j)=\sqrt{1-cc_I^2(i,j)}$ cc_I correlation coefficient of intensity

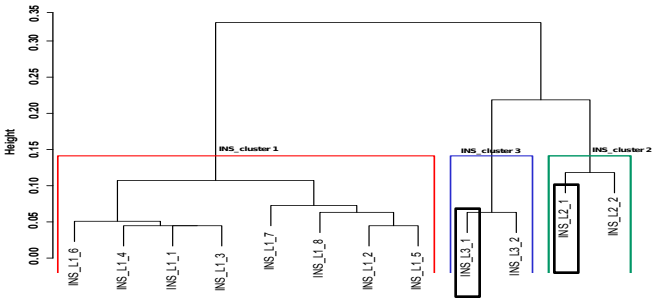
$$L(A,.B) = \frac{1}{N_A \times N_B}\sum_{p=1}^{N_A}\sum_{q=1}^{N_B} d(i_p, j_q)$$ Average Linkage

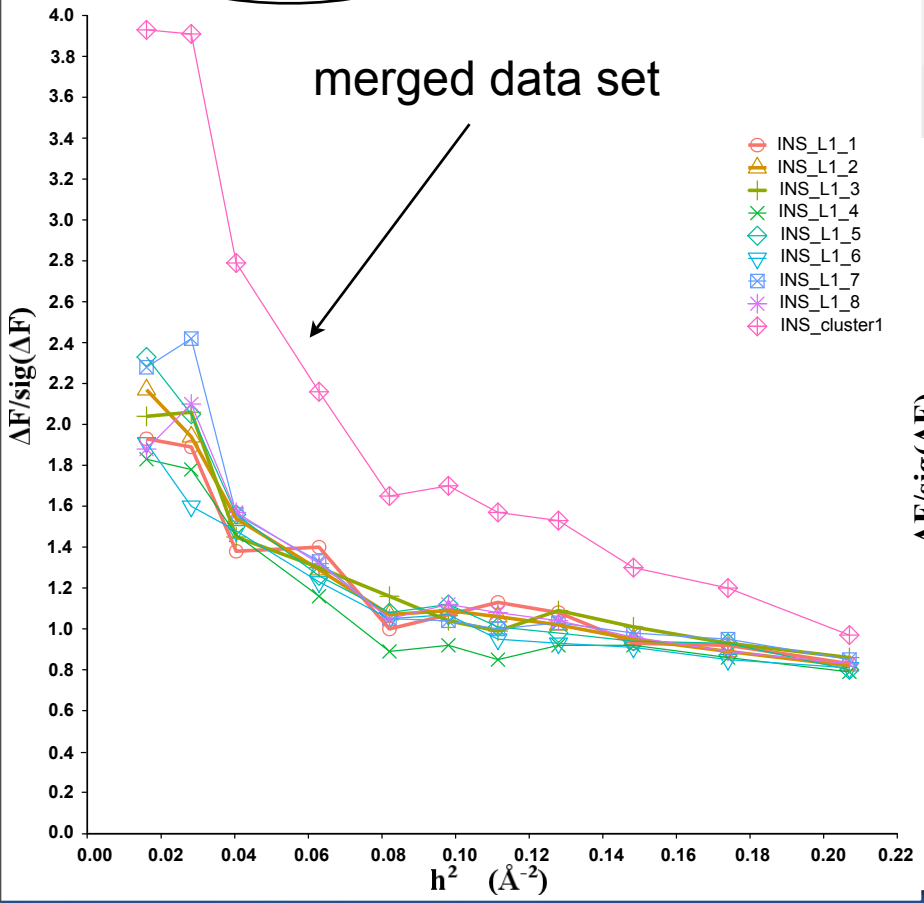
Insulin

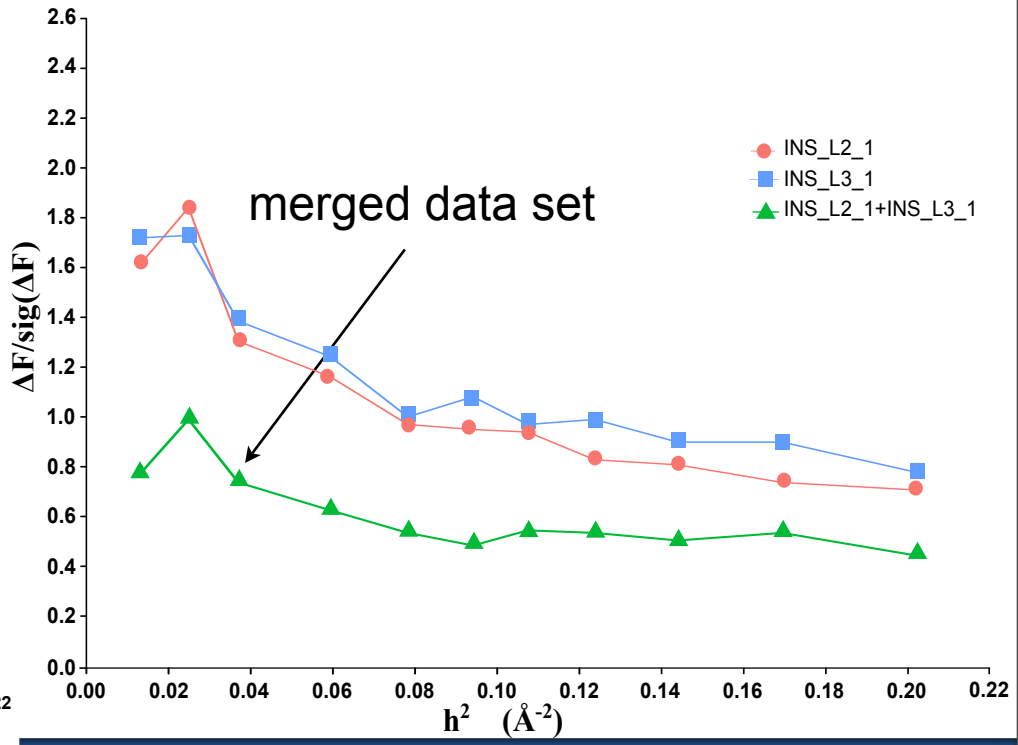

Height
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
INS_cluster 1
INS_cluster 3
INS_cluster 2
INS_L1_6
INS_L1_4
INS_L1_1
INS_L1_3
INS_L1_7
INS_L1_8
INS_L1_2
INS_L1_5
INS_L3_1
INS_L3_2
INS_L2_1
INS_L2_2

Insulin

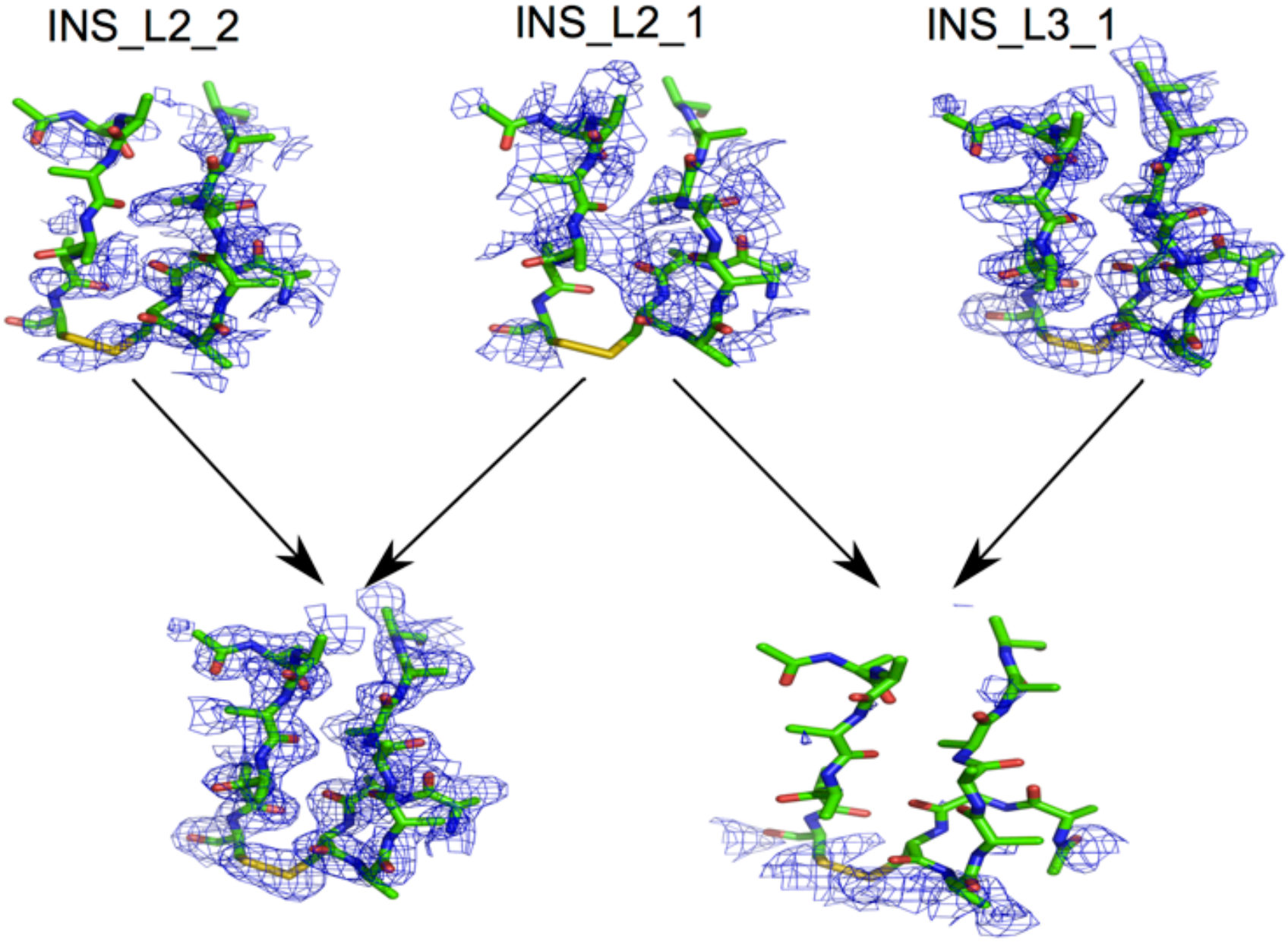

Height
INS_cluster 1
INS_cluster 3
INS_cluster 2
INS_L1_6
INS_L1_4
INS_L1_1
INS_L1_3
INS_L1_7
INS_L1_8
INS_L1_2
INS_L1_5
INS_L3_1
INS_L3_2
INS_L2_1
INS_L2_2


Height
INS_cluster 1
INS_cluster 3
INS_cluster 2
INS_L1_6
INS_L1_4
INS_L1_1
INS_L1_3
INS_L1_7
INS_L1_8
INS_L1_2
INS_L1_5
INS_L3_1
INS_L3_2
INS_L2_1
INS_L2_2

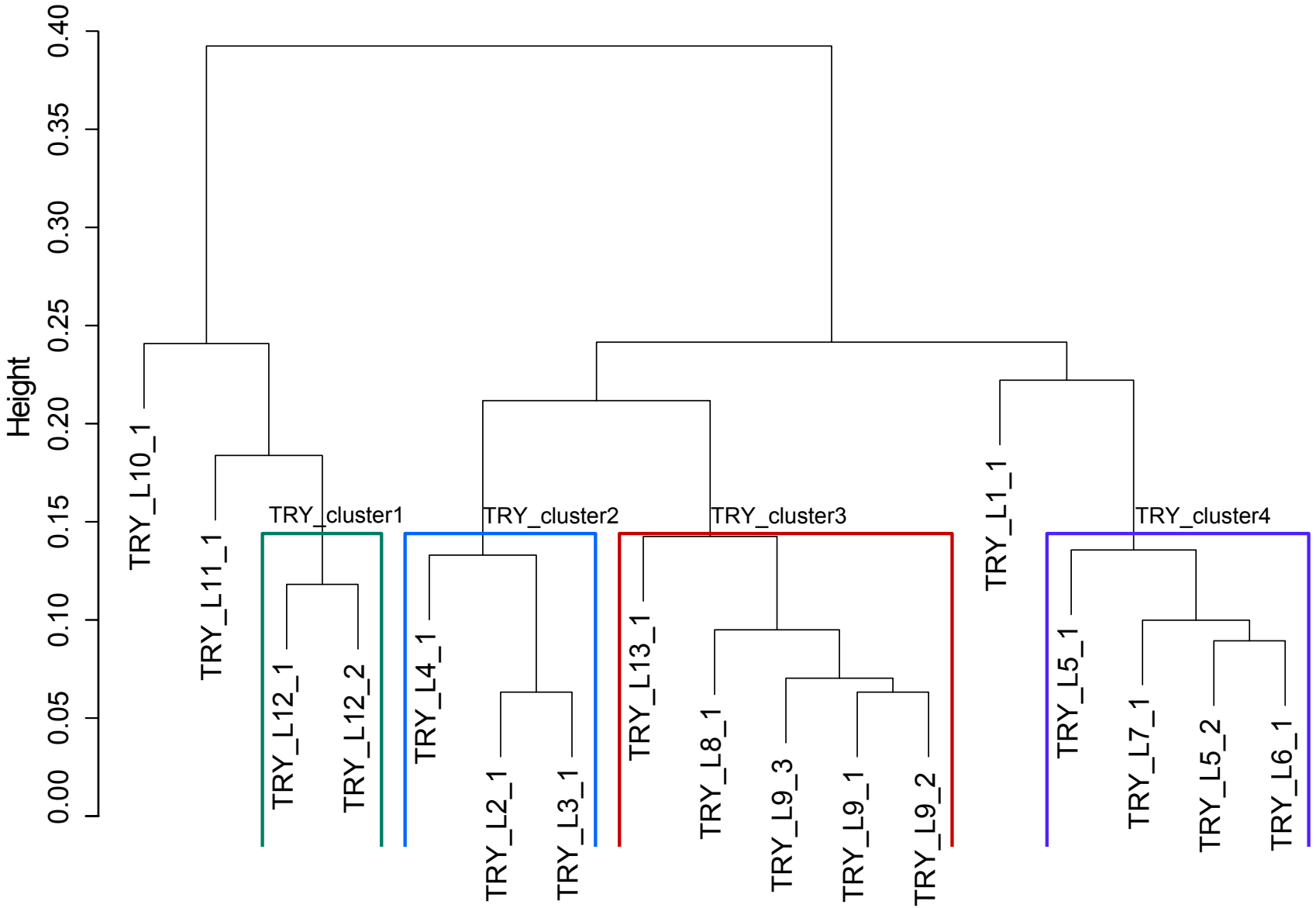

merged data set
INS_L1_1
INS_L1_2
INS_L1_3
INS_L1_4
INS_L1_5
INS_L1_6
INS_L1_7
INS_L1_8
INS_cluster1
ΔF/sig(ΔF)
h² (Å⁻²)


merged data set
INS_L2_1
INS_L3_1
INS_L2_1+INS_L3_1
ΔF/sig(ΔF)
h² (Å⁻²)

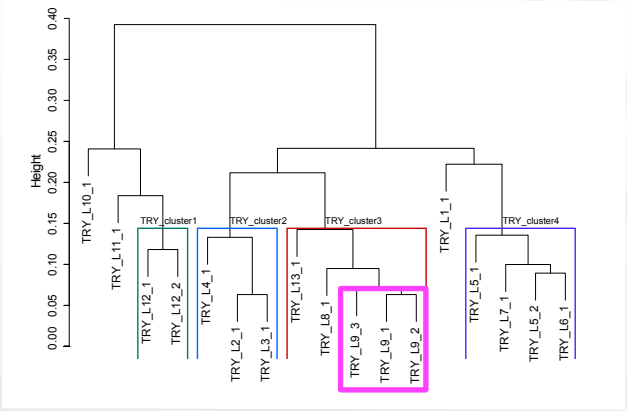
Insulin



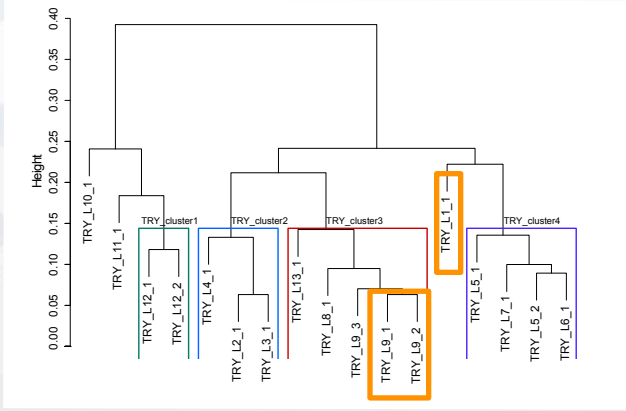
TRYPSIN

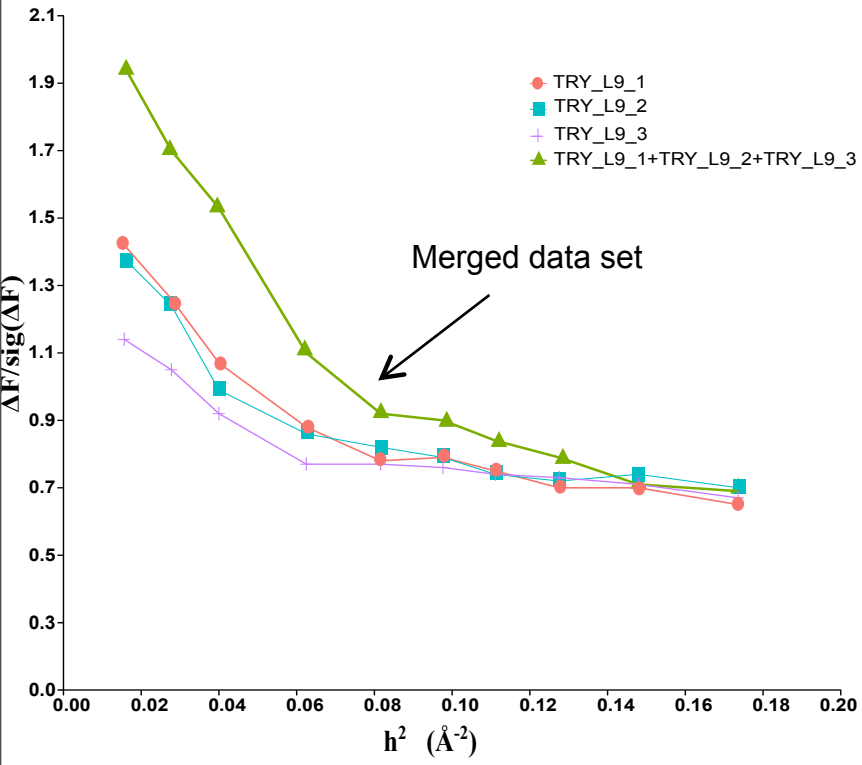

Height
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
TRY_L10_1
TRY_L11_1
TRY_L12_1
TRY_L12_2
TRY_cluster1
TRY_L4_1
TRY_L2_1
TRY_L3_1
TRY_cluster2
TRY_L13_1
TRY_L8_1
TRY_L9_3
TRY_L9_1
TRY_L9_2
TRY_cluster3
TRY_L1_1
TRY_L5_1
TRY_L7_1
TRY_L5_2
TRY_L6_1
TRY_cluster4

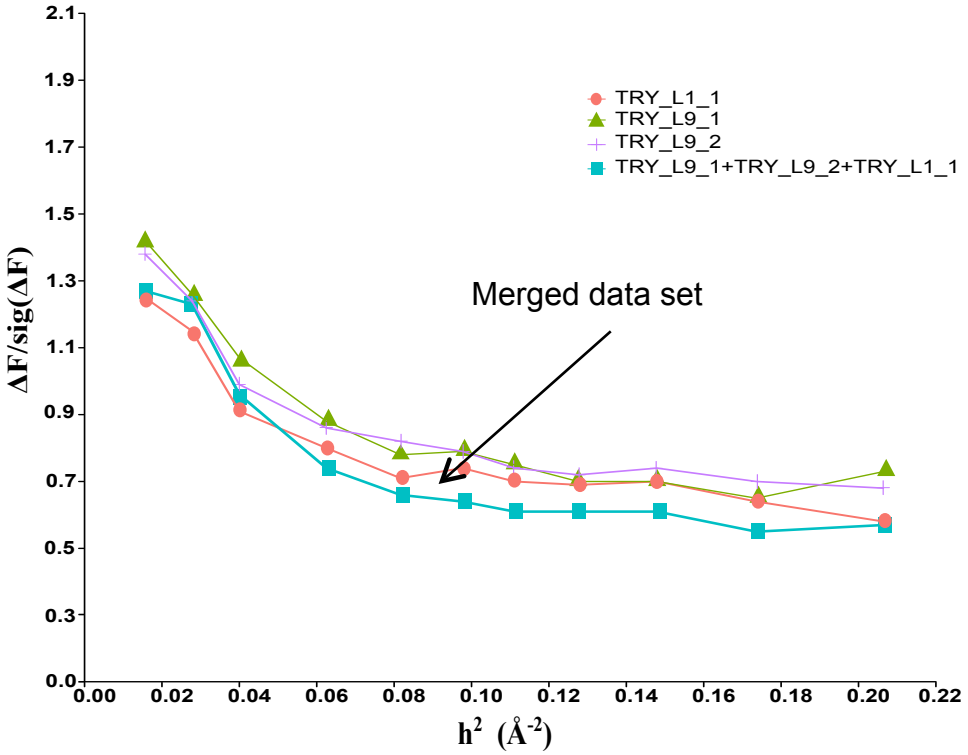
Data sets from the same cluster



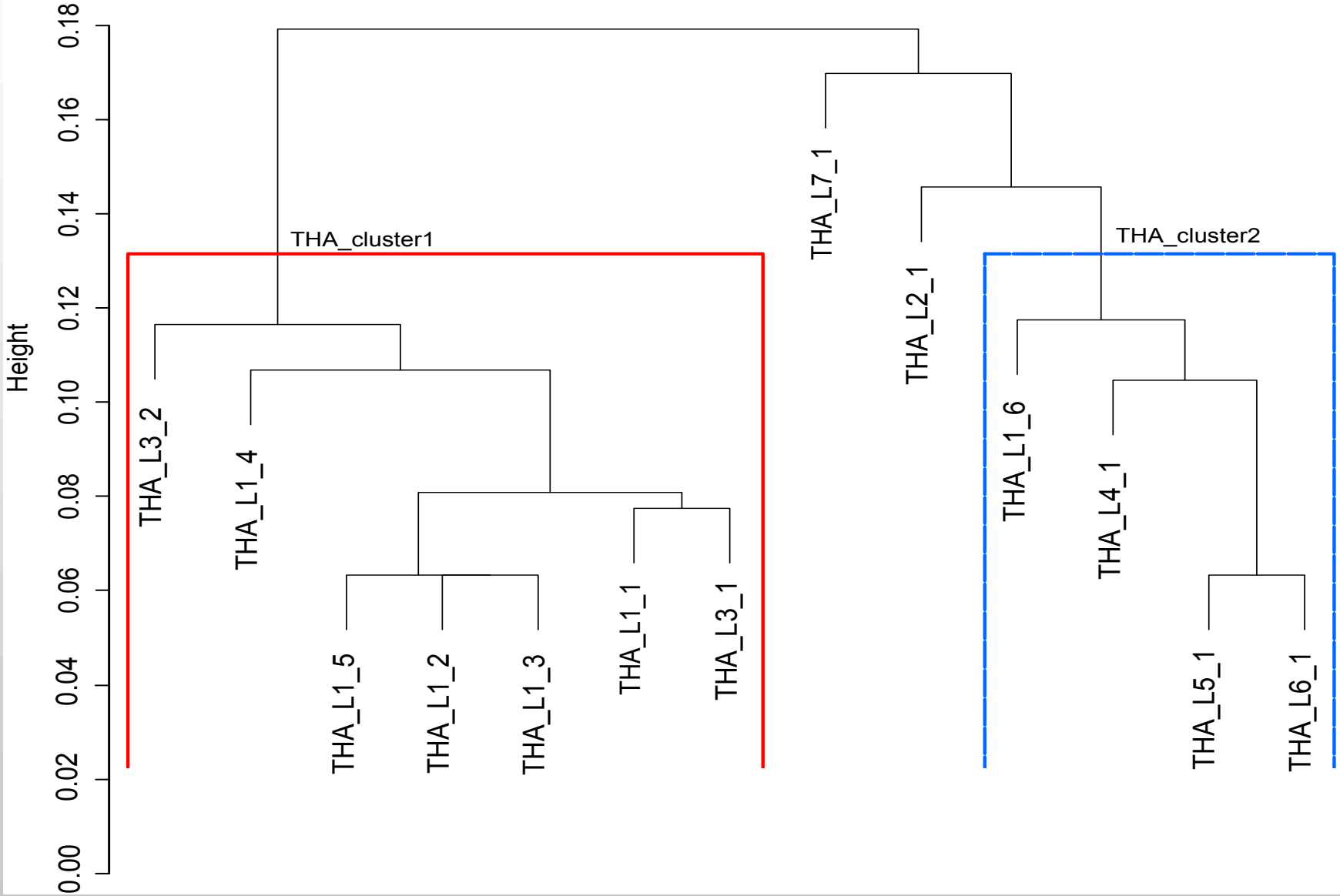
Data sets from different clusters



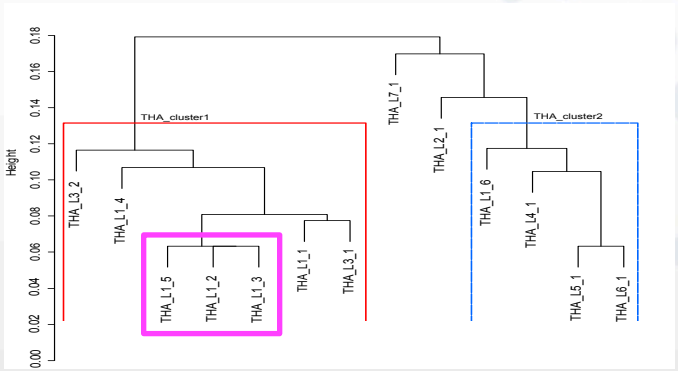




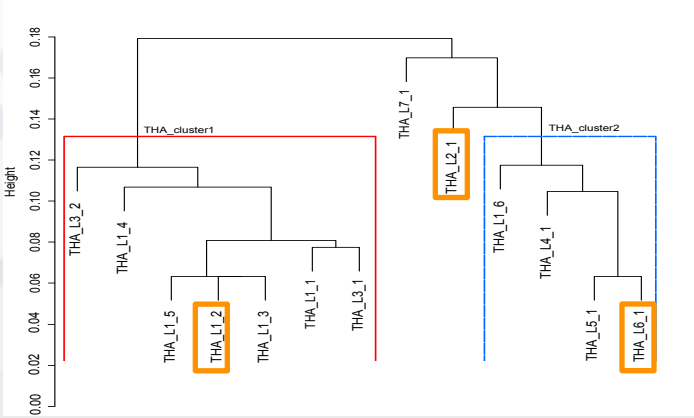
THAUMATIN

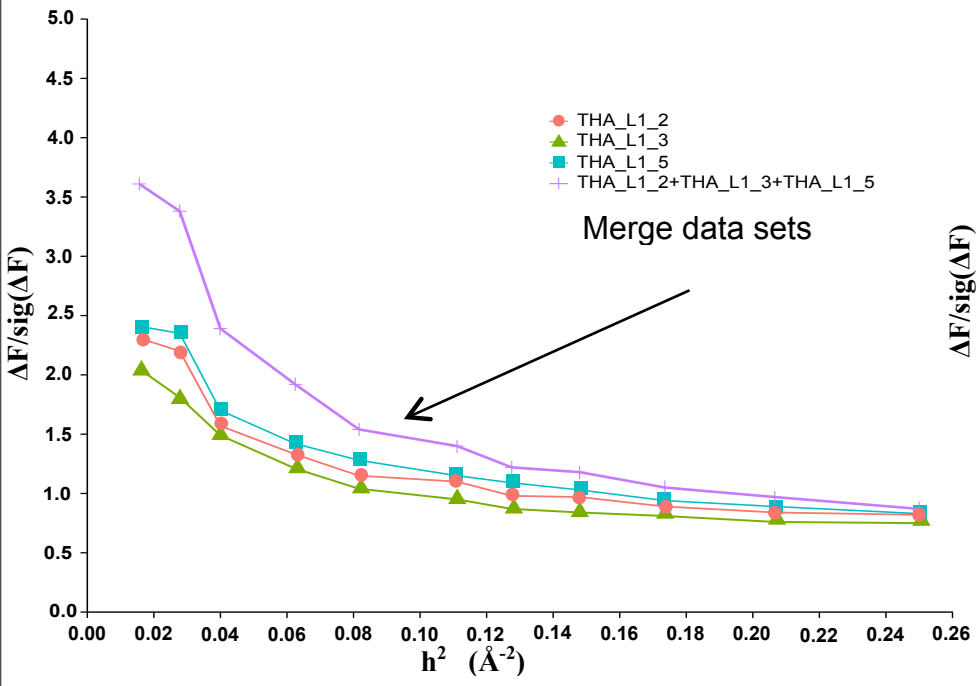

Height
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
THA_cluster1
THA_cluster2
THA_L3_2
THA_L1_4
THA_L1_5
THA_L1_2
THA_L1_3
THA_L1_1
THA_L3_1
THA_L7_1
THA_L2_1
THA_L1_6
THA_L4_1
THA_L5_1
THA_L6_1

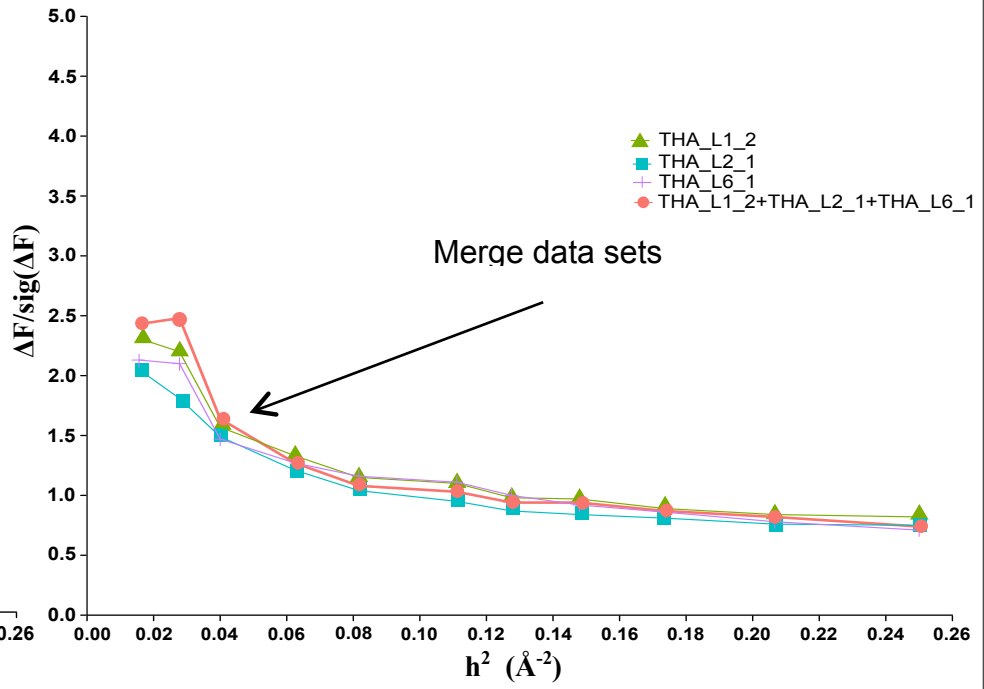
Data sets from the same cluster



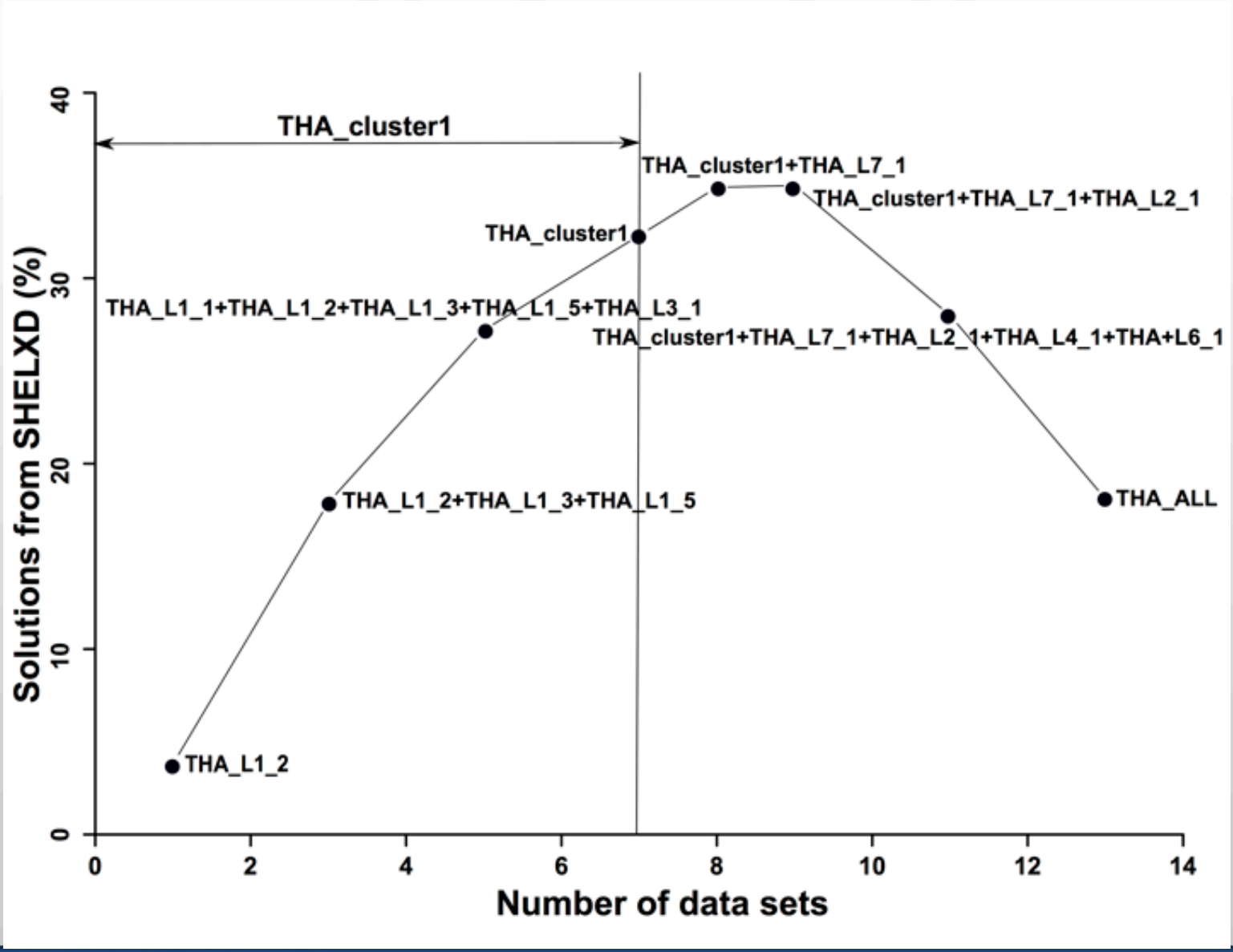
Data sets from different clusters



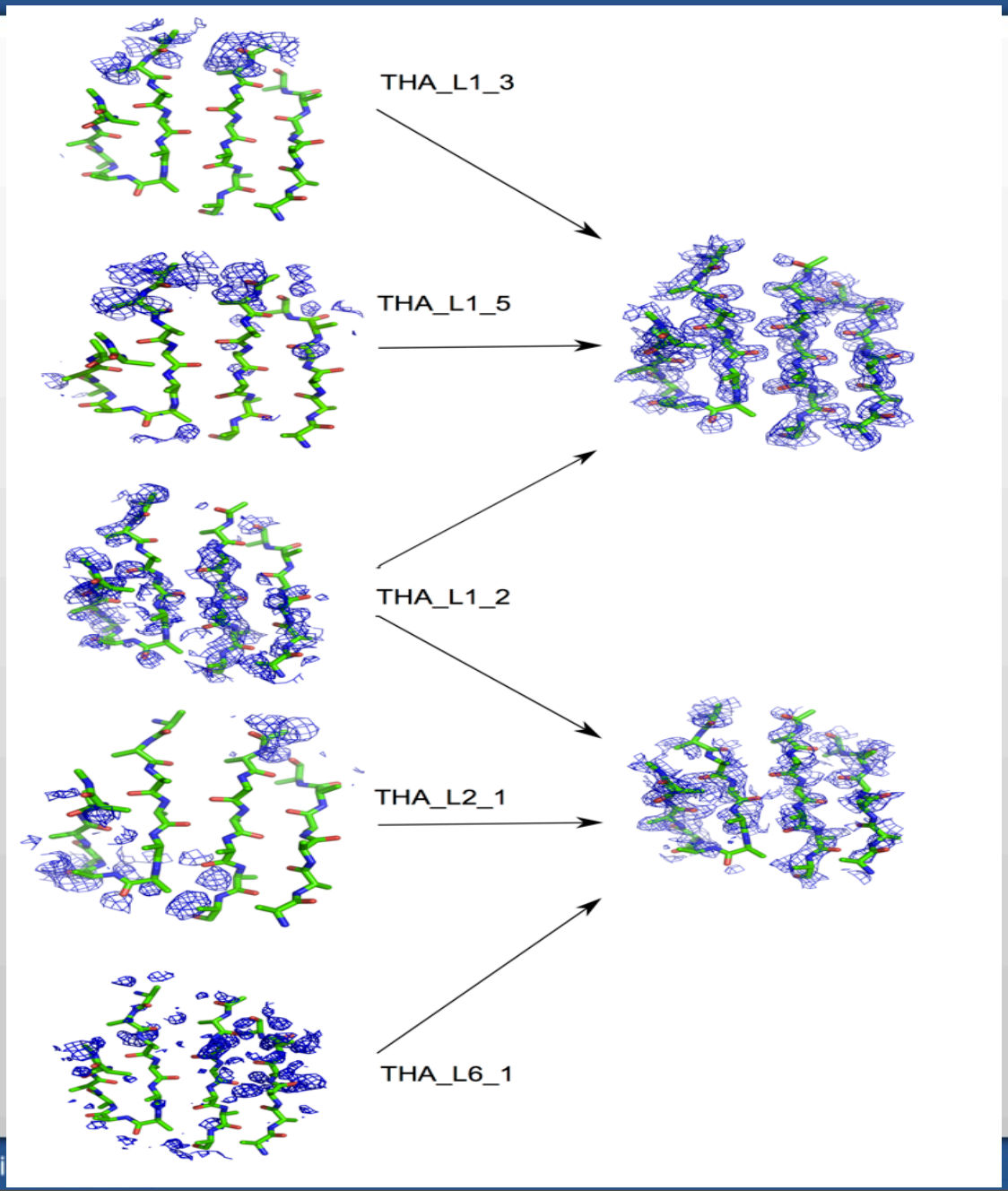




THAUMATIN


THA_cluster1
THA_cluster1+THA_L7_1
THA_cluster1+THA_L7_1+THA_L2_1
THA_cluster1
THA_L1_1+THA_L1_2+THA_L1_3+THA_L1_5+THA_L3_1
THA_cluster1+THA_L7_1+THA_L2_1+THA_L4_1+THA+L6_1
THA_L1_2+THA_L1_3+THA_L1_5
THA_ALL
THA_L1_2
Solutions from SHELXD (%)
0
10
20
30
40
0
2
4
6
8
10
12
14
Number of data sets

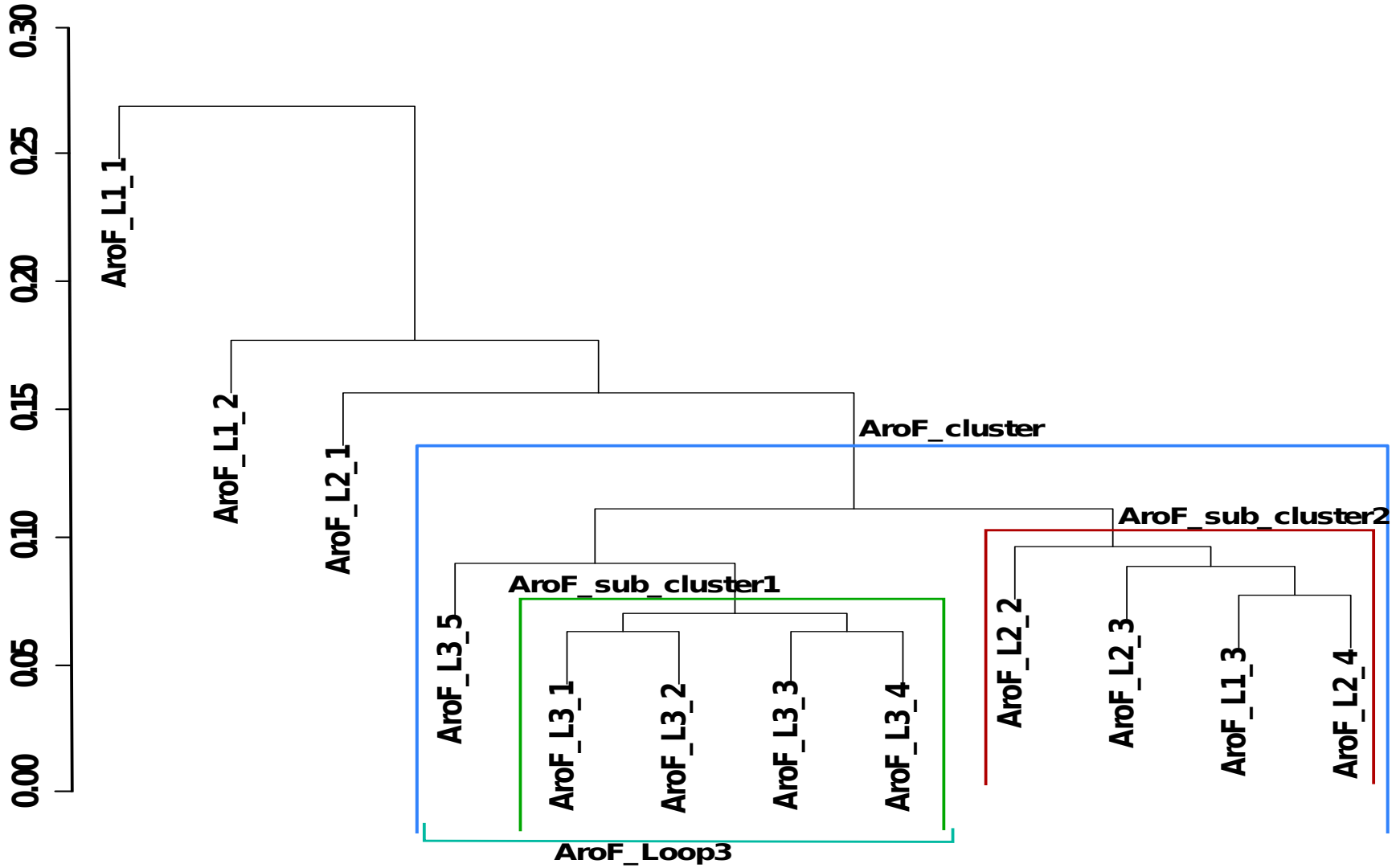
THAUMATIN



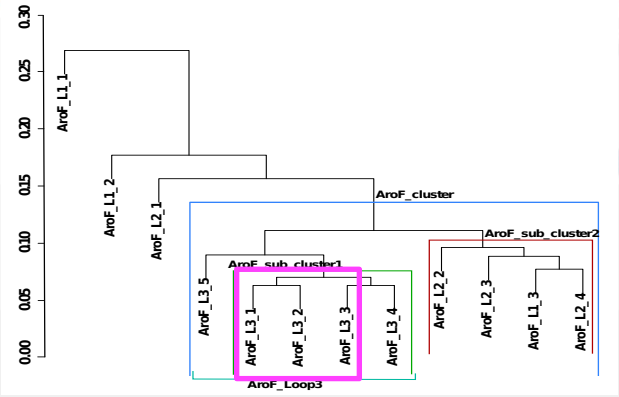
AroF

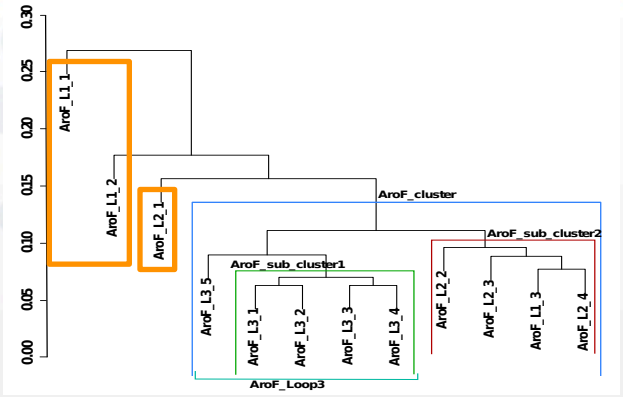
11 Selenomethionines

$P6_422$

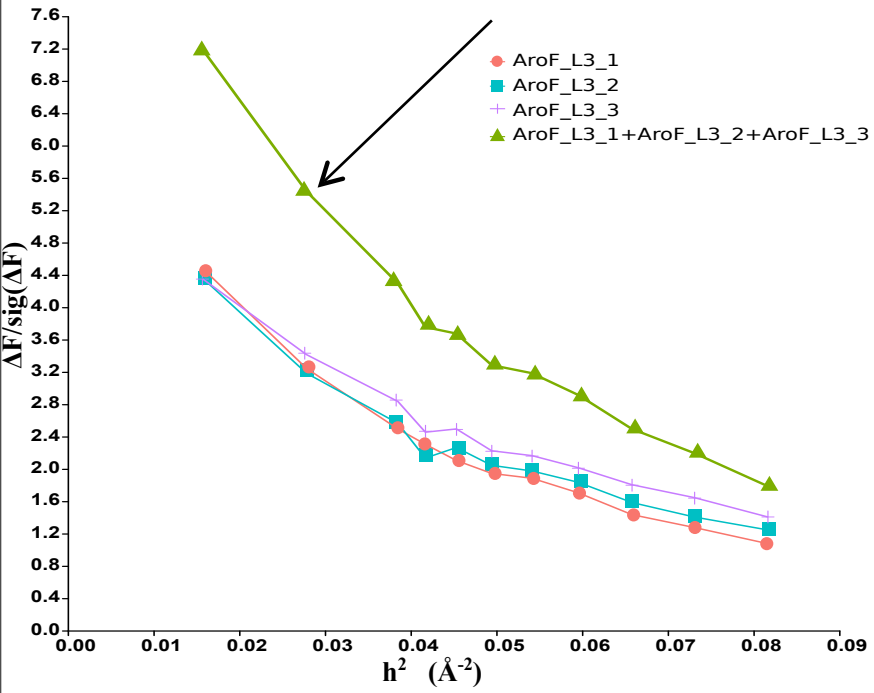


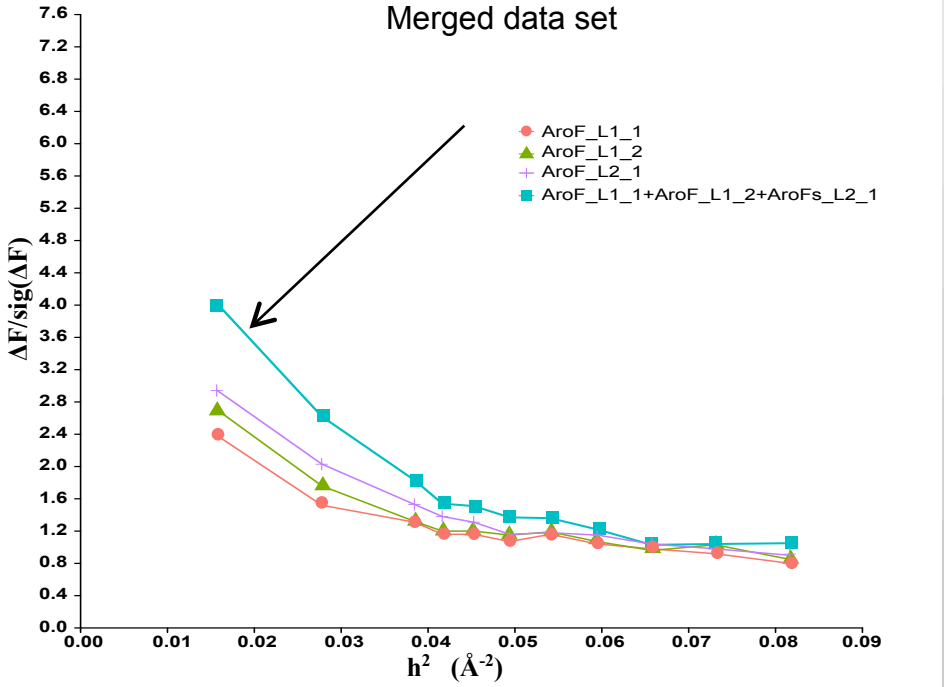
AroF



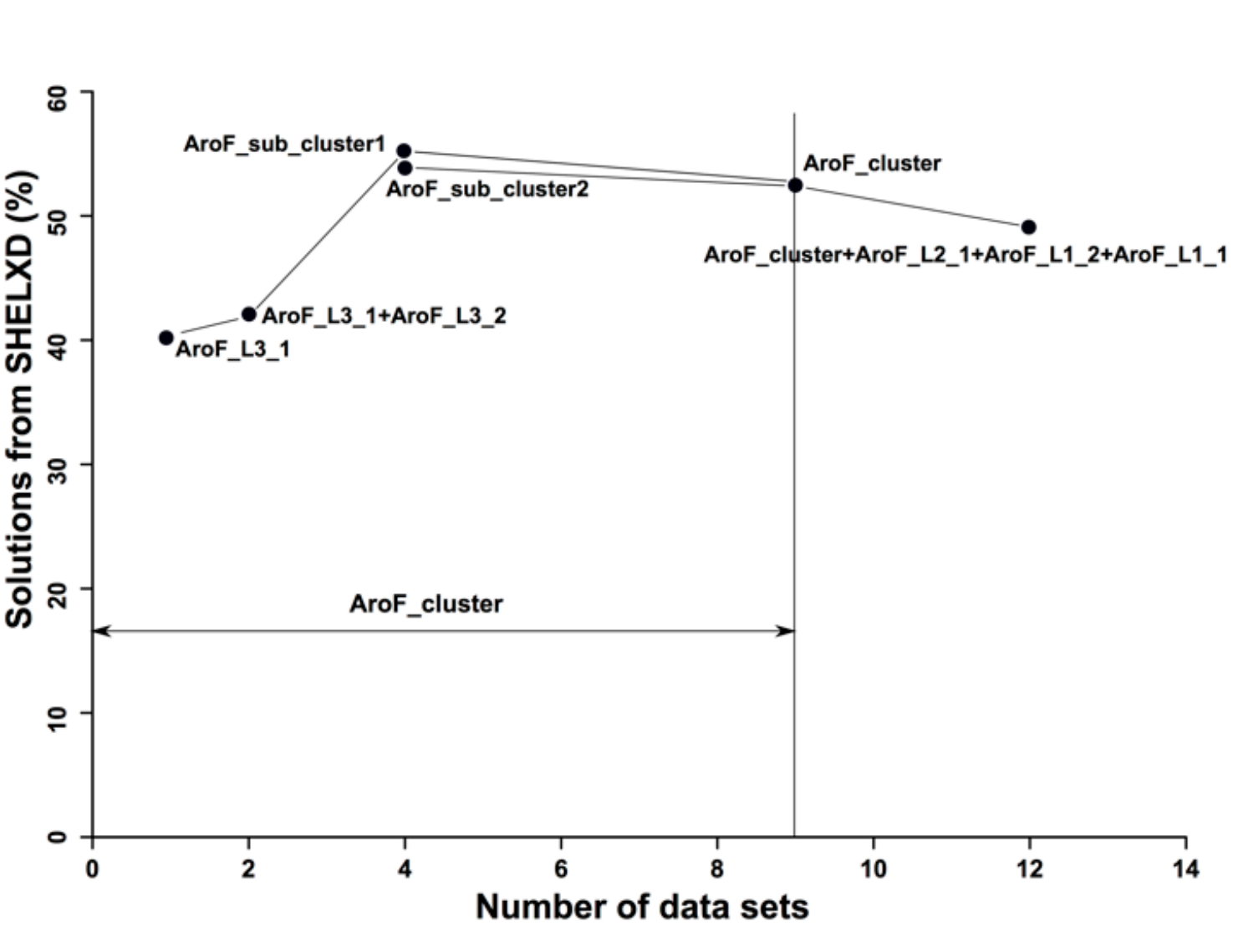


Merged data set

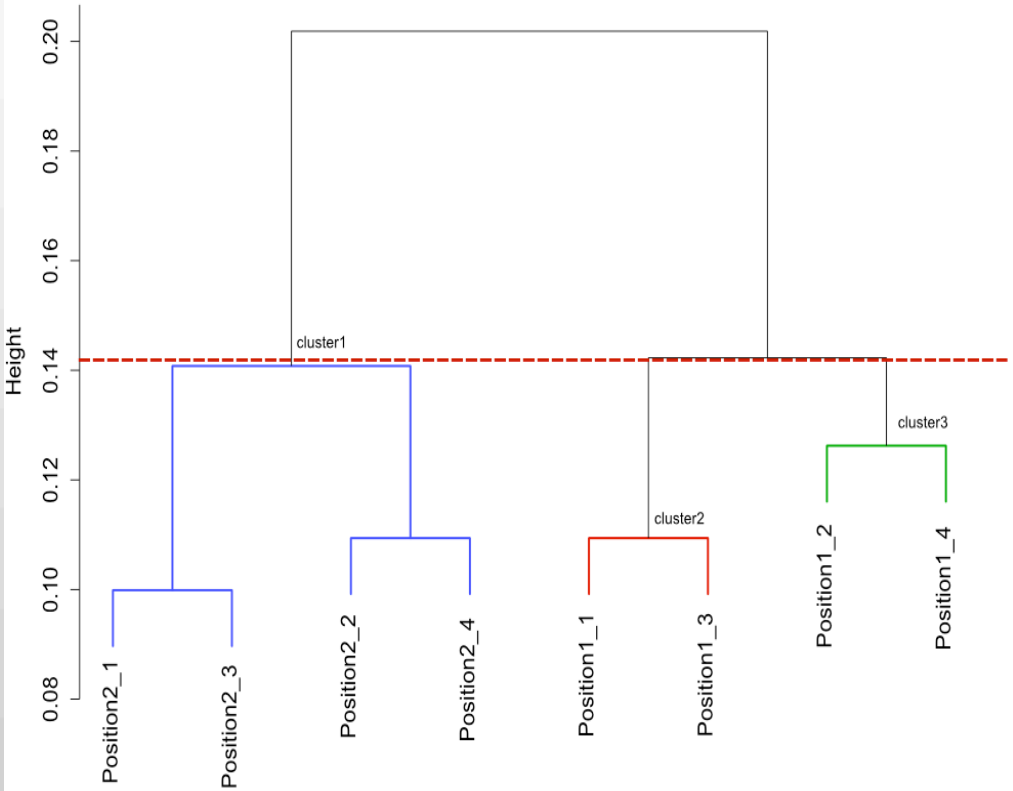




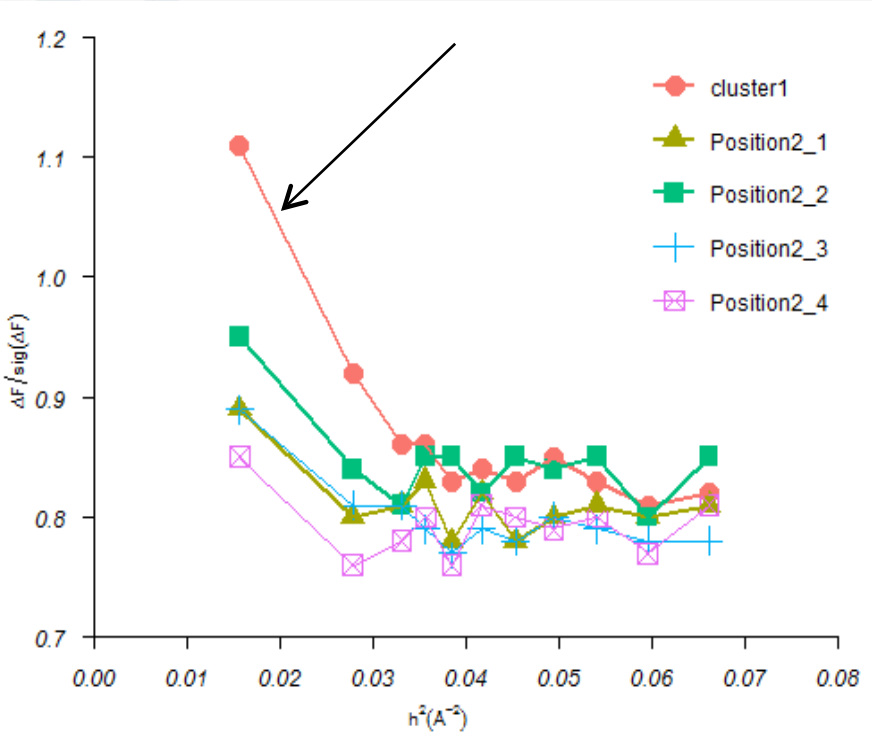
AroF


AroF_sub_cluster1
AroF_sub_cluster2
AroF_cluster
AroF_cluster+AroF_L2_1+AroF_L1_2+AroF_L1_1
AroF_L3_1+AroF_L3_2
AroF_L3_1
AroF_cluster
Solutions from SHELXD (%)
0
10
20
30
40
50
60
0
2
4
6
8
10
12
14
Number of data sets

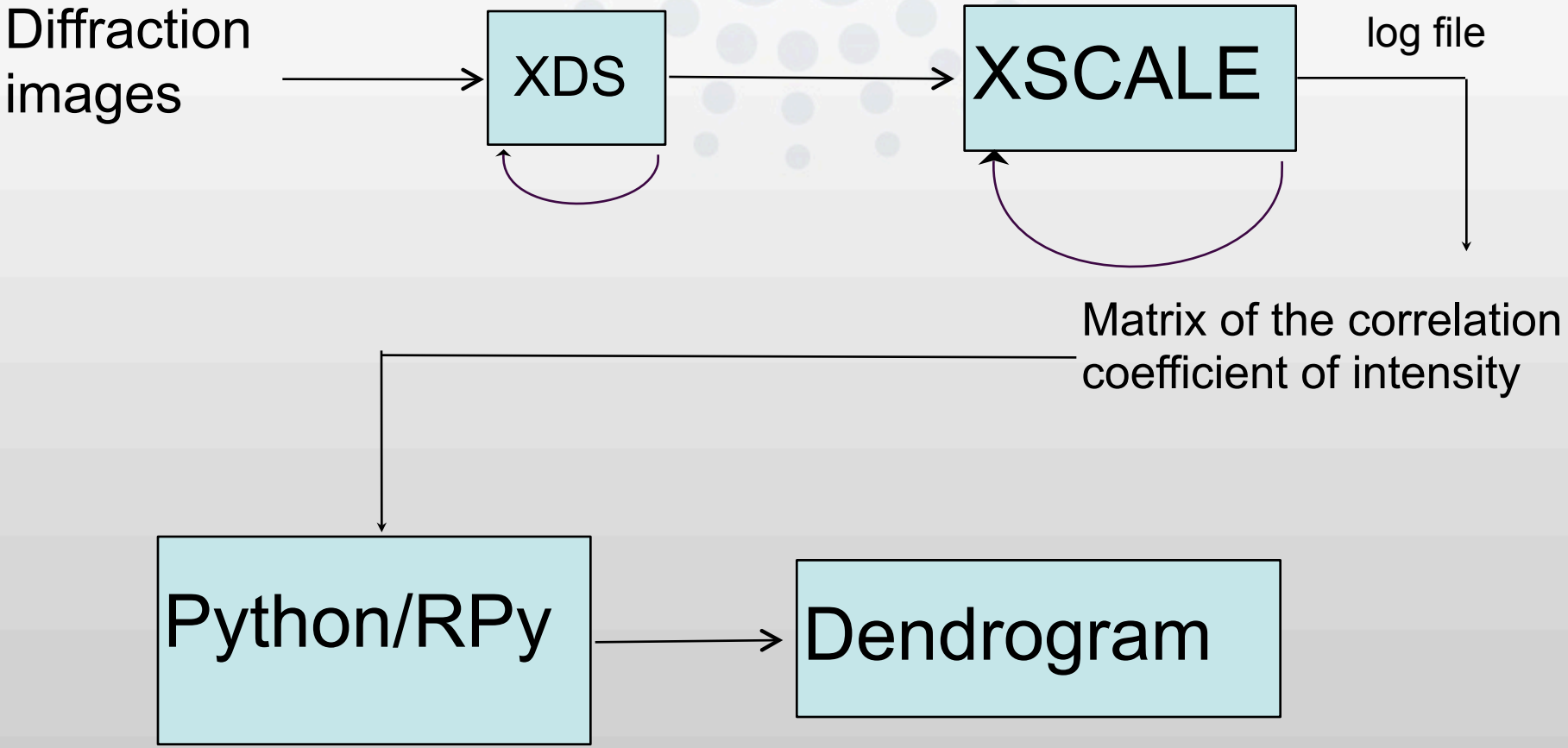
2 select position of the same crystal

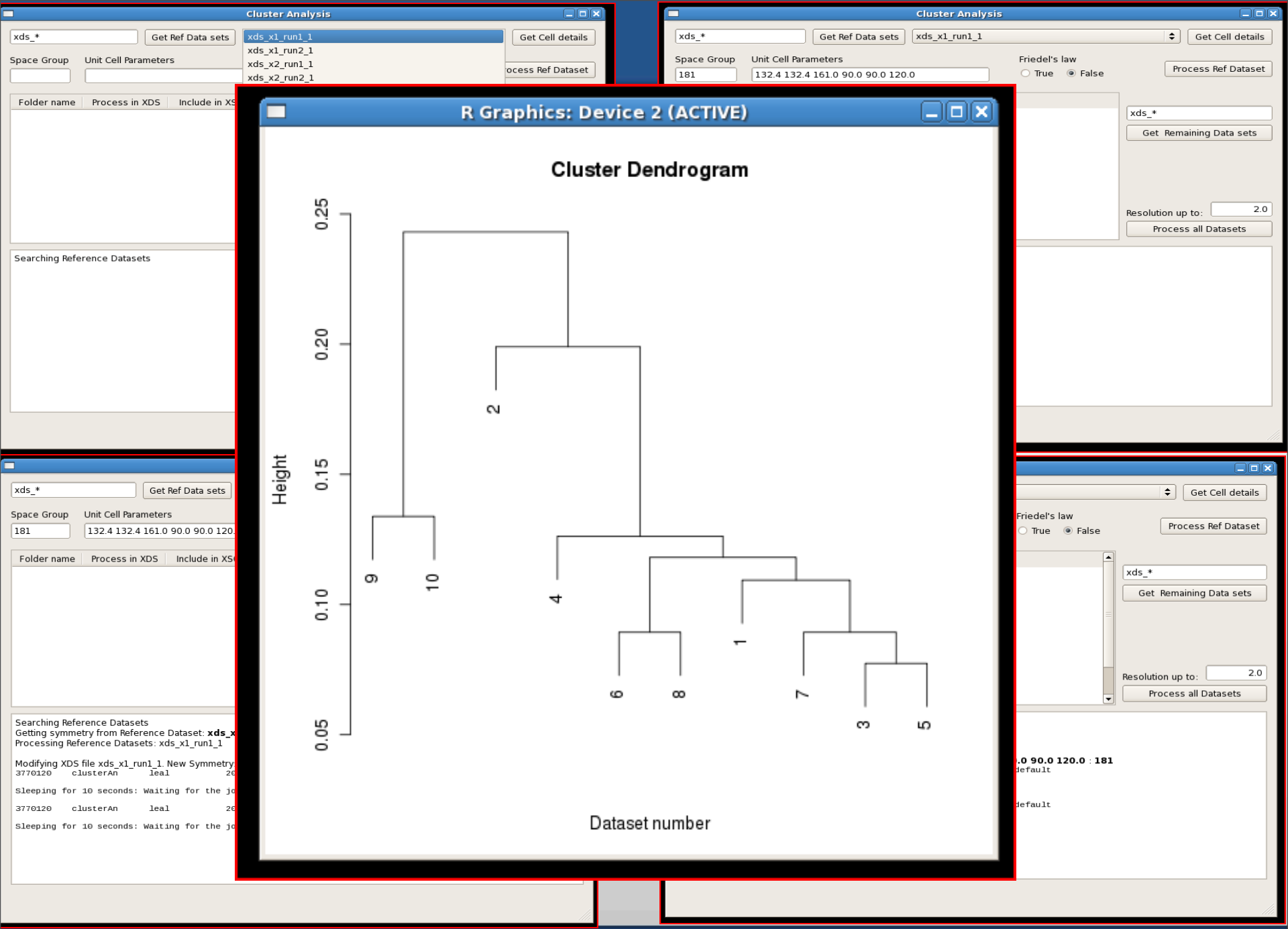


Merged data set



Software implementation





Courtesy of Ricardo Leal

CONCLUSION

1. Using pair correlation coefficients, the proteins crystals under study can be clearly separated into clusters.
2. Data collection statistics, for anomalous x-ray diffraction, can be improved by merging isomorphic data sets from frozen crystals. In the case of non-isomorphism the quality of the merged data set will be worse than the individual data sets.
3. The software implementation of the method is currently ready to use.

Acknowledgements

ESRF

Alexander (Sasha) Popov
Sean McSweeney
Ricardo Leal
Gordon Leonard
Ulrike Kapp
Matthew Bowler

Universidad Carlos III de Madrid
Departamento de Matemáticas

Giovanna Miritello

Thank you for your attention